

A PTAS for the Minimum Consensus Clustering Problem with a Fixed Number of Clusters

Paola Bonizzoni* Gianluca Della Vedova† Riccardo Dondi‡

July 10, 2009

Abstract

The Consensus Clustering problem has been introduced as an effective way to analyze the results of different microarray experiments [5, 6]. The problem consists of looking for a partition that best summarizes a set of input partitions (each corresponding to a different microarray experiment) under a simple and intuitive cost function. The problem admits polynomial time algorithms on two input partitions, but is APX-hard on three input partitions. We investigate the restriction of Consensus Clustering when the output partition is required to contain at most k sets, giving a polynomial time approximation scheme (PTAS) while proving the NP-hardness of this restriction.

1 Introduction

Microarray data analysis is a fundamental task in studying genes. Indeed, microarray experiments provide measures of gene expression levels under certain experimental conditions, showing that groups of genes have a similar behavior under certain conditions. However, even slightly different experimental conditions may result in significantly different expression data. These gene expression patterns are useful to understand the relations among genes and could provide information useful for the construction of genetic networks. Nowadays the use of microarrays has become widespread and

*Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy, bonizzoni@disco.unimib.it

†Dipartimento di Statistica, Università degli Studi di Milano-Bicocca, Milano, Italy, gianluca.dellavedova@unimib.it

‡Dipartimento di Scienze dei Linguaggi, della Comunicazione e degli Studi Culturali, Università degli Studi di Bergamo, Bergamo, Italy, riccardo.dondi@unibg.it

sufficiently cheap to justify running a large battery of experiments under similar, albeit not identical, conditions. The integration of the results is therefore the final computational step needed to obtain a meaningful interpretation of the data.

In [5, 6] a clustering approach to the integration of different experimental microarray experimental data was introduced. In the proposed approach, called **CONSENSUS CLUSTERING**, the genes are represented by elements of a universe set. The experimental data under certain experimental condition, are represented as a partition of the universe set, where a set represents elements (genes) that have similar expression level in the experiment. The proposed approach then computes the consensus of the partitions given by a collection of gene expression data, since integrating different experimental data is potentially more informative than the individual experimental data. More precisely, **CONSENSUS CLUSTERING** asks for a partition of the universe set that better summarizes a set of input partitions on the same universe. The **CONSENSUS CLUSTERING** problem has been studied extensively in the literature and its NP-hardness over general instances is well-known [9, 11].

The minimization version of **CONSENSUS CLUSTERING**, called **MINIMUM CONSENSUS CLUSTERING**, admits a $\frac{3}{2}$ -approximation algorithm [1] as well as a number of heuristics based on cutting-plane [8] and simulated annealing [6]. In the latter paper, it was observed that the problem is trivially solvable for instances of at most two partitions, while an open question, as recently recalled [1], is the computational complexity of the problem (for both minimization and maximization versions) on k input partitions, for any constant $k > 2$. The question has been settled in [3] by showing that **MINIMUM CONSENSUS CLUSTERING** is APX-hard even on instances with three input partitions, hence making hopeless the search for a polynomial time algorithm. In this paper we will focus on the restriction of the problem where the *desired* consensus partition has at most k sets, with k a constant.

A problem closely related to **MINIMUM CONSENSUS CLUSTERING** is **MINIMUM CORRELATION CLUSTERING**. In **MINIMUM CORRELATION CLUSTERING**, given a complete graph where each edge is associated with a label in $\{+, -\}$, the goal is to compute a partition of the vertices of the graph so that the number of co-clustered vertices joined by $-$ edges and the number of vertices joined by $+$ edges and not co-clustered is minimized. The restriction of **MINIMUM CORRELATION CLUSTERING** where the output partition has at most k sets, is NP-hard but admits a PTAS [7]. We will extend the analysis of [7] by showing that the analogous restriction **MINIMUM CONSENSUS CLUSTERING** admits a PTAS, while being NP-hard.

Notice that **MINIMUM CORRELATION CLUSTERING** and **MINIMUM CON-**

SENSUS CLUSTERING are not comparable, since the input graph in MINIMUM CORRELATION CLUSTERING is unweighted, while the input graph of MINIMUM CORRELATION CLUSTERING is weighted. On the other hand, it is quite immediate to notice that there are unweighted graphs that are not an instance of MINIMUM CONSENSUS CLUSTERING.

2 The problem

We will tackle the CONSENSUS CLUSTERING problem, in its minimization version. Two elements of the universe set are *co-clustered* in a partition π if they belong to the same set of π .

Definition 2.1. Let V be a universe set and let π_1, π_2 be two partitions of V . Let $d(\pi_1, \pi_2)$ denote the *symmetric difference distance* defined as the number of pairs of elements co-clustered in exactly one of π_1 and π_2 . Let $s(\pi_1, \pi_2)$ denote the *similarity measure* defined as the number of pairs of elements co-clustered in both partitions plus the number of pairs of elements not co-clustered in both partitions π_1 and π_2 .

Given two elements i, j of the universe set V and a set $\Pi = \{\pi_1, \dots, \pi_l\}$ of partitions of V , we denote by $s_\Pi(i, j)$ (or simply $s(i, j)$ whenever Π is known from the context) and the distance $d_\Pi(i, j)$ (or simply $d(i, j)$) respectively, the number of partitions of Π in which i, j are co-clustered and are not co-clustered. Clearly, for each pair (i, j) , $d_\Pi(i, j) + s_\Pi(i, j) = l$, that is the number of partitions. When Π consists of 2 partitions π_1 and π_2 , we denote by $d(\pi_1, \pi_2)$ the quantity $\sum_{i < j} d_{\{\pi_1, \pi_2\}}(i, j)$.

We are now able to formally introduce the problem we will study in this paper, MINIMUM CONSENSUS CLUSTERING when the output partition is required to have at most k sets (denoted by K-MIN-CC): we are given a set $\Pi = \{\pi_1, \pi_2, \dots, \pi_l\}$ of partitions over universe V and we want to find a partition π of V , such that π has at most k sets and π minimizes $d(\pi, \Pi) = \sum_{i=1}^l d(\pi, \pi_i)$, that is the cost of solution π . In what follows, we denote by K-MIN-CC(l) the restriction of the K-MIN-CC problem where the input consists of exactly l partitions of V .

The MINIMUM CONSENSUS CLUSTERING is closely related to the MINIMUM CORRELATION CLUSTERING [2], where we are given a labeled complete graph, with each edge labeled by either $+$ or $-$ and the goal is to compute a partition C_1, C_2, \dots, C_k of the vertex set so that the number of $+$ edges cut by the partition and the number of $-$ edges inside a same set C_i is minimized. Several variants of the correlation clustering have been introduced [4, 10, 1].

An instance of MINIMUM CONSENSUS CLUSTERING can be represented with a labeled complete graph $G = (V, E)$, where each edge $(v, w) \in E$ is labeled by $s_{\Pi}(v, w)$. In Section 3 we assume that the instance of K-MIN-CC(l) is precisely this graph representation of MINIMUM CONSENSUS CLUSTERING.

3 The PTAS

In this section we will show that the K-MIN-CC admits a PTAS, that is for any $\epsilon > 0$ a polynomial time approximation algorithm with a guaranteed $1 + \epsilon$ ratio between the costs of the approximate solution and the optimal solution. Let $G = (V, E)$ be the complete graph instance of K-MIN-CC.

The MinDisAg algorithm of [7] for MINIMUM CORRELATION CLUSTERING can be restated to solve K-MIN-CC and is reported here as Alg. 1. Let us detail the idea behind MinDisAg [7] and how it can be generalized. First of all, some “small” instances are solved by a brute force approach, namely when only one set must be computed or when the number n of input elements is polynomial in k (the number of desired output sets). In fact, there are at most k^n possible partitions of V , and k^n is a constant whenever n is polynomial in k .

The algorithm starts by randomly sampling a subset S of V . If the sample is not too large (i.e. $O(\log n)$), then it is possible to compute all partitions of S in polynomial time. Since the steps that the algorithm performs for each partition require polynomial time, the whole algorithm has polynomial time complexity.

The algorithm extends each partition of S to a partition of V . Since the number of partitions of S is polynomial, we can restrict our attention only to the partition \tilde{S} that fully agrees on S with the overall optimal solution \mathcal{D} . On that specific partition, extending \tilde{S} to a partition of V introduces only a few errors.

More precisely, the algorithm applies a greedy procedure to extend \tilde{S} : it assigns independently each element x of $V \setminus S$ to the cluster of \tilde{S} that minimizes the total cost of all pairs made of x and an element of S .

This procedure computes a clustering of V into sets that can be distinguished into large and small, depending on the fact that a set is smaller or larger than a certain threshold. The large sets are retained, while all small sets are merged together obtaining a new universe set which is in turn recursively fed to the algorithm (only this time requiring a smaller error ratio and obtaining a partition with fewer sets.)

We remember that l denotes the number of input partitions, and k denotes the number of sets in the output partition. Given a partition P of V , the cost of P is denoted by $\text{cost}(P)$. Let ϵ' be equal to $\frac{\epsilon}{128 \cdot 20^2 k^4}$ (i.e. ϵ' is a constant depending only on ϵ and the number of sets in the output partition). We distinguish two cases: the optimum is at most $\epsilon' n^2$ or at least $\epsilon' n^2$. In the latter case we exploit the fact that it is possible to solve the problem in polynomial time and with a guaranteed *additive* error $\epsilon \epsilon' n^2$, where n is the number of elements in the universe, for any constant $\epsilon > 0$ (see [3] for details). Then the approximation ratio is at most $\frac{\epsilon \epsilon' n^2 + \epsilon' n^2}{\epsilon' n^2} = 1 + \epsilon$, that is the algorithm in [3] computes the required approximate solution. Therefore, in the following we only have to investigate the case when the optimal solution has a cost at most $\epsilon' n^2$.

We define $t = \frac{2560000k^4}{\epsilon^2} \log n$ as the size of the sample set S , $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_k\}$ as the optimal solution (whose cost is denoted by γn^2). Let \tilde{S} be the partition $\{X \cap S : X \in \mathcal{D}\}$, that is the restriction of \mathcal{D} to the set S . We recall that we will mainly focus on the iteration of steps 6–21 where such \tilde{S} is extended to a partition of the universe set V . Let \mathcal{A} be a partition of a set $A \subseteq V$, and let x be an element of V . Then $N^{\mathcal{A}}(x)$ is the set of all elements of A different from x and co-clustered with x in \mathcal{A} . Given an element $u \in V$, define $\text{val}_i^{\mathcal{A}}(u)$:

$$\text{val}_i^{\mathcal{A}}(u) = \frac{1}{|A \setminus \{u\}|} (|\{x \in N^{\mathcal{A}}(u) \wedge s(x, u) = i\}| + |\{x \notin N^{\mathcal{A}}(u) \wedge d(x, u) = i\}|).$$

Informally $\text{val}_i^{\mathcal{A}}(u)$ is the fraction of pairs consisting of u and an element of A that may give a contribution $l - i$ to the cost of the solution. Moreover we define $\text{val}^{\mathcal{A}}(u)$ as

$$\text{val}^{\mathcal{A}}(u) = \frac{\sum_{x \in N^{\mathcal{A}}(u)} s(x, u) + \sum_{x \notin N^{\mathcal{A}}(u)} d(x, u)}{l|A \setminus \{u\}|}.$$

Informally $\text{val}^{\mathcal{A}}(u)$ is the fraction of input pairs containing u on which \mathcal{A} agrees. Notice that $\text{val}^{\mathcal{A}}(u) = \frac{1}{l} \sum_{i=1}^l i \cdot \text{val}_i^{\mathcal{A}}(u)$. Let \mathcal{A} be a partition of the set A , then $\mathcal{A}(u, i)$ is the partition obtained from \mathcal{A} moving the element u to the set A_i (notice that u may not belong to A). Given an integer j , with $1 \leq j \leq l$, define $\text{pval}^{\mathcal{A}}(u, i) = \text{val}^{\mathcal{A}(u, i)}(u)$ and $\text{pval}_j^{\mathcal{A}}(u, i) = \text{val}_j^{\mathcal{A}(u, i)}(u)$. Finally we introduce the notion of β -good partition, which is a good approximation of the optimal partition. Let X be a subset of V , \mathcal{A} be a partition of A and $\beta = \frac{\epsilon}{128 \cdot 20^2 k^4}$. Then \mathcal{A} is β -good if for each $u \in V$, $0 \leq j \leq l$ and $1 \leq i \leq k$, then

$$|\text{pval}_j^{\mathcal{A}}(u, i) - \text{pval}_j^{\mathcal{D}}(u, i)| \leq \beta.$$

Algorithm 1: MinDisAg(k, ϵ)

Input: A set Π of partitions of V

Output: A k -clustering of the graph, i.e. a partition of V into at most k sets V_1, \dots, V_k

```
1 if  $k = 1$  then
2   Return the obvious 1-clustering;
3 if  $n \leq 16k^2$  then
4   Return the optimal  $k$ -clustering, obtained by exhaustive search;
5 ClusMax  $\leftarrow$  the result of the PTAS for Max Consensus Clustering [3]
   with accuracy  $\bar{\epsilon}(\epsilon, k)$ ;
6 Pick a sample  $S \subseteq V$  by drawing  $|S| = \frac{500 \log n}{\beta^2}$  elements uniformly at
   random with replacement;
7  $m \leftarrow \infty$ ;
8 foreach each partition  $\bar{S}$  of  $S$ ,  $\bar{S} = \{S_1, \dots, S_k\}$  do
9   Initialize the clusters  $C_i \leftarrow S_i$  for  $1 \leq i \leq k$ ;
10  for each  $u \in V \setminus S$  do
11     $j_u \leftarrow \arg \min_i \{cost(\bar{S} \setminus S_i \cup (S_i \cup \{u\}))\}$ ;
    /*  $j_u$  maximizes  $pval^{\bar{S}}(u, j_u)$  */
    /*  $val^{\bar{S}}(u) \leftarrow pval^{\bar{S}}(u, j_u)$  */
12    Add  $u$  to the set  $C_{j_u}$ ;
    /* Compute the set of large and small clusters */
13     $Large \leftarrow \{j | 1 \leq j \leq k, |C_j| \geq \frac{n}{2k}\}$ ;
14     $Small \leftarrow \{1, \dots, k\} \setminus Large$ ;
15     $l \leftarrow |Large|$  and  $s \leftarrow k - l = |Small|$ ;
16     $W \leftarrow \bigcup_{j \in Small} C_j$ ;
17     $\Pi' \leftarrow$  the restriction of the partitions in  $\Pi$  to the new universe set
     $W$ ;
18    Recursively call MinDisAg on the partitions  $\Pi'$  and with
    arguments  $(s, \epsilon/3)$ . Denote by  $W'_1, W'_2, \dots, W'_s$  the result;
19     $\mathcal{C} \leftarrow \{C_1, \dots, C_l, W'_1, \dots, W'_s\}$ ;
20    if  $cost(\mathcal{C}) < m$  then
21       $m \leftarrow cost(\mathcal{C})$ ;
22      ClusMin  $\leftarrow \mathcal{C}$ ;
23 Return the better of the two clusterings ClusMax and ClusMin;
```

3.1 Analysis of the Algorithm

Notice that the main contribution of this section lies in Lemma 3.1 which is a stronger version of a result in [7]; in that paper the notion of $pval^A(u, i)$ is sufficient because the problem studied is unweighted. In our paper we study a problem where each pair of elements can have a cost that is an integer between 0 and l , therefore we need a definition of $pval_j^A(u, i)$, with a new parameter j expressing the number of input partitions where two elements are either co-clustered or not co-clustered. Indeed our definition of β -goodness requires that a certain inequality holds for values of j that are integers between 0 and l , while in [7] j can – implicitly – only take 0 or 1 as value.

Recall that we denote by \tilde{S} the restriction of \mathcal{D} to the sample set S . The following lemma proves that \tilde{S} is, with high probability, a good sample of the optimal solution.

Lemma 3.1. *The partition \tilde{S} is β -good with probability at least $1 - O(\frac{1}{\sqrt{n}})$.*

Proof. Let v be an element of S and let u be an element of V . Let $p(v, i, j)$ be a variable equal to 1 if and only if $v \in N_{\mathcal{D}(u, i)}(u)$ and $s(v, u) = j$ or $v \notin N_{\mathcal{D}(u, i)}(u)$ and $d(v, u) = j$. Pose $p(v, i, j) = 0$ otherwise.

By construction of $p(v, i, j)$ and $pval_j^{\mathcal{D}}(u, i)$, the probability $Pr[p(v, i, j) = 1] = pval_j^{\mathcal{D}}(v, i)$, as the set S is sampled randomly from V . Also notice that $pval_j^{\tilde{S}}(v, i) = val_j^{\tilde{S}(v, i)}(v) = \frac{1}{|S \setminus \{u\}|} \left(|\{x \in N_{\tilde{S}(v, i)}(v) \wedge s(x, v) = j\}| + |\{x \notin N_{\tilde{S}(v, i)}(v) \wedge d(x, v) = j\}| \right) = \frac{1}{|S \setminus \{u\}|} \sum_{v \in S \setminus \{u\}} p(v, i, j)$, as the latter equality is an immediate consequence of the definition of $p(v, i, j)$.

The Hoeffding bound states that, given some causal variables X_i such that $Pr[X_i = 1] = p$ (and $X_i = 0$ otherwise), then $Pr[|X_a - \frac{1}{m} \sum_{a=1}^m X_a| > \beta] \leq 2e^{-2m\beta^2}$. In our case the causal variable X_a are $p(v, i, j)$, and the sum is over all elements $v \in S \setminus \{u\}$, therefore the inequality becomes $Pr[|p(v, i, j) - \frac{1}{|S \setminus \{u\}|} \sum_{v \in S \setminus \{u\}} p(v, i, j)| > \beta] \leq 2e^{-2(|S \setminus \{u\}|)\beta^2} \leq 2e^{-2t\beta^2}$. By the previous arguments, the inequality can be rewritten as $Pr[|pval_j^{\mathcal{D}}(u, i) - pval_j^{\tilde{S}}(u, i)| > \beta] \leq 2e^{-2t\beta^2}$, which gives an upper bound on the probability that any element $u \in V$ does not satisfy the requirements of an β -good set.

Applying a union bound we obtain that the probability of having at least one of the t elements not satisfying the requirements is at most $2te^{-2t\beta^2}$. Since $|S| = \frac{500 \log n}{\beta^2}$, the partition \tilde{S} is β -good with probability at least $1 - 2 \frac{500 \log n}{\beta^2} e^{-1000 \log n} = 1 - 2 \frac{500 \cdot 160^2 \cdot 20^2 k^4 l^2 \log n}{\beta \epsilon^2} \frac{1}{n^{1000}}$, which is trivially larger than $1 - \frac{c}{\sqrt{n}}$ for some constant c . \square

We will now provide some simple generalizations of the Lemmas in [7], omitting the proofs as they are straightforward extensions of those in [7]. Just as in [7], we will assume that the sample S is β -good, for some constant β , and we will focus on the iteration of the algorithm for the partition \bar{S} of S that agrees with the optimal partition D . We will denote by C_1, \dots, C_k the sets in ClusMin at the end of such iteration.

Lemma 3.2 (Lemma 4.3 in [7]). *Let $u \in V \setminus S$ with $u \in \mathcal{D}_s$ (that is the s -th set of the optimal solution), and $u \in C_r$ for $r \neq s$ (that is u is misplaced by the algorithm). Then $pval_j^{\mathcal{D}}(u, r) \geq pval_j^{\mathcal{D}}(u, s) - 2\beta = val_j^{\mathcal{D}}(u) - 2\beta$ for each $0 \leq j \leq k$.*

Recall that l is the number of input partitions, define T_{low} as the set $\{u \in V : val^{\mathcal{D}}(u) \leq 1 - \frac{1}{20k^2}\}$, and let us call *bad* all elements in T_{low} and *good* all elements that are not in T_{low} . As each element u in T_{low} contributes to the cost of a solution of κ -MIN-CC(l) for at least $\frac{1}{2}l(n-1)(1 - val^{\mathcal{D}}(u)) \leq \frac{1}{40k^2}l(n-1)$, a simple counting argument allows us to prove that there are at most $\frac{80\gamma nk^2}{l}$ bad elements.

For clarity's sake, we split Lemma 4.4 in [7] into two separate statements, where the first statement (Lemma 3.3) is actually proved in the first part of the proof of Lemma 4.4 in [7], while the second statement corresponds to Lemma 4.4 in [7]. Those technical results show that (i) our algorithm clusters almost optimally all good elements and (ii) all good elements in *Large* are optimally clustered, pending a condition on various parameters that will be proved at the end of the section (for the definition of *Large* and *Small* see Algorithm 1). More precisely, Lemma 3.3 states that misplaced good elements must belong to some small sets (which in turn implies that the majority of good elements must be optimally clustered).

Lemma 3.3. *Let u be an element in $C_i \setminus T_{low}$ but not in $\mathcal{D}_i \setminus T_{low}$. Then $u \in \mathcal{D}_j$, for some $j \neq i$, and $|\mathcal{D}_i| \leq 2(\frac{1}{20k^2} + \beta)n + 1$.*

Proof. The proof is the same as in [7], except for the observation that, by our definition of $pval$ and since each pair of elements involving u is correctly co-clustered when u is in either \mathcal{D}_i or \mathcal{D}_j , $pval^{\mathcal{D}}(u, j) + pval^{\mathcal{D}}(u, i) \leq 2 - \frac{l(|\mathcal{D}_i| + |\mathcal{D}_j| - 1)}{l(n-1)}$. \square

Lemma 3.4. *Let i be an element in *Large*. If $\frac{n}{2k} - \gamma n^2 \frac{40k^2}{l(n-1)} > 2(k + 1)((\frac{1}{20k^2} + \beta)n + 1)$ and $2(\frac{1}{20k^2} + \beta)n + k < \frac{n}{2k} - \frac{80\gamma nk^2}{l}$ then $C_i \setminus T_{low} = \mathcal{D}_i \setminus T_{low}$.*

Proof. Let $x \in V \setminus T_{low}$. W.l.o.g. we can assume that $x \in C_1 \setminus T_{low}$ and $x \in \mathcal{D}_1 \cup T_{low}$. First we will prove that $C_1 \setminus T_{low} \subseteq \mathcal{D}_1 \setminus T_{low}$. Assume to the contrary that there exists a $y \in C_1$, $y \notin \mathcal{D}_1, T_{low}$, therefore (w.l.o.g.) $y \in \mathcal{D}_2$. By Lemma 3.3, and since there are at most k sets in \mathcal{D} , then $|C_1 \setminus (\mathcal{D}_1 \cup T_{low})| \leq 2(\frac{1}{20k} + \beta k)n + k$.

Since $C_1 \setminus (\mathcal{D}_1 \cup T_{low}) = (C_1 \setminus \mathcal{D}_1) \setminus T_{low} = (C_1 \setminus T_{low}) \setminus \mathcal{D}_1$ then $|\mathcal{D}_1| \geq |C_1 \setminus T_{low}| - |C_1 \setminus (\mathcal{D}_1 \cup T_{low})| \geq |C_1| - |T_{low}| - |C_1 \setminus (\mathcal{D}_1 \cup T_{low})| \geq \frac{n}{2k} - \gamma n^2 \frac{40k^2}{l(n-1)} - 2(\frac{1}{20k} + \beta k)n + k$. But $\frac{n}{2k} - \gamma n^2 \frac{40k^2}{l(n-1)} - 2(\frac{1}{20k} + \beta k)n + k > 2(\frac{1}{20k^2} + \beta)n + 1$, which contradicts $|\mathcal{D}_1| \leq 2(\frac{1}{20k^2} + \beta)n + 1$. In fact $\frac{n}{2k} - \gamma n^2 \frac{40k^2}{l(n-1)} - 2(\frac{1}{20k} + \beta k)n + k > 2(\frac{1}{20k^2} + \beta)n + 1$ can be rewritten as $\frac{n}{2k} - \gamma n^2 \frac{40k^2}{l(n-1)} > 2(k + 1)(\frac{1}{20k^2} + \beta)n + 1$.

Now we know that $C_1 \setminus T_{low} \subseteq \mathcal{D}_1 \setminus T_{low}$ and we would like to prove that $C_1 \setminus T_{low} \supseteq \mathcal{D}_1 \setminus T_{low}$, along the same lines as for the first part. Assume to the contrary that there exists a $y \in \mathcal{D}_1$, $y \notin C_1, T_{low}$, therefore (w.l.o.g.) $y \in C_2$. Again by Lemma 3.3, both \mathcal{D}_1 and \mathcal{D}_2 have at most $2(\frac{1}{20k^2} + \beta)n + 1$ elements. Notice that $C_1 \setminus T_{low} \subseteq \mathcal{D}_1$, since $C_1 \setminus T_{low} \subseteq \mathcal{D}_1 \setminus T_{low}$, moreover C_1 is large, therefore $|C_1| \geq \frac{n}{2k}$. By the value of $|T_{low}|$, $2(\frac{1}{20k^2} + \beta)n + k \geq \frac{n}{2k} - \frac{80\gamma nk^2}{l}$ which does not hold by hypothesis. \square

Now we are able to show that there is a solution where some sets are exactly the large sets in ClusMin and whose cost is not much larger than the optimum. This fact justifies the recursive step of the algorithm. The condition under which the lemma holds will be proved at the end of the section.

Lemma 3.5. *If $l(n-1)|T_{low}| \left(2\beta + \frac{|T_{low}|}{l(n-1)}\right) \leq \frac{\epsilon}{2}\gamma n^2$, then there exists a solution $F = \{F_1, \dots, F_k\}$ such that the cost of F is at most $\gamma n^2(1 + \epsilon/2)$ and $F_i = C_i$ for each i in Large.*

Proof. Let F be the solution consisting of all large sets in ClusMin and where all remaining elements are partitioned as in \mathcal{D} . Clearly the only pairs of elements that might not be partitioned in F as in ClusMin are the ones containing at least one element of T_{low} , by Lemma 3.4. By the definition of val , $cost(F) - cost(\mathcal{D}) \leq l(n-1) \sum_{x \in T_{low}} (val^{\mathcal{D}}(x) - val^F(x))$.

We have to consider two different cases, depending on the fact that $x \in T_{low}$ belongs to sets C_i , \mathcal{D}_i for a certain i , or not. In the first case w.l.o.g. x is in both C_1 and \mathcal{D}_1 the set of pairs that are different in ClusMin and in \mathcal{D} , are only pairs of the form (x, y) where $y \in T_{low}$, which in turn implies that $val^F(x) \geq val^{\mathcal{D}}(x) - \frac{|T_{low}|}{l(n-1)}$. In the second case we

can assume w.l.o.g. that $x \in C_1$ and $x \in \mathcal{D}_2$. Applying Lemma 3.2 we know that $pval^{\mathcal{D}}(x, y) \geq val^{\mathcal{D}}(x) - 2\beta$. Also notice that in $\mathcal{D}(x, 1)$ and F , the element x belong to the same set therefore, just as for the first case, $val^F(x) \geq val^{\mathcal{D}(x, 2)}(x) - \frac{|T_{low}|}{l(n-1)}$, but $val^{\mathcal{D}(x, 2)}(x) = pval^{\mathcal{D}}(x, 2)$. Combining all inequalities we obtain $val^F(x) \geq pval^{\mathcal{D}}(x, 2) - \frac{|T_{low}|}{l(n-1)} \geq val^{\mathcal{D}}(x) - 2\beta - \frac{|T_{low}|}{l(n-1)}$, where the last inequality comes from Lemma 3.2. In both cases we can say that $val^F(x) \geq pval^{\mathcal{D}}(x, 2) - \frac{|T_{low}|}{l(n-1)} \geq val^{\mathcal{D}}(x) - 2\beta - \frac{|T_{low}|}{l(n-1)}$. An immediate consequence is that $cost(F) - cost(\mathcal{D})$ is at most $l(n-1) \sum_{x \in T_{low}} (val^F(x) - val^{\mathcal{D}}(x)) \leq l(n-1)|T_{low}| \left(2\beta + \frac{|T_{low}|}{l(n-1)}\right)$. The claim follows since $l(n-1)|T_{low}| \left(2\beta + \frac{|T_{low}|}{l(n-1)}\right) \leq \gamma n^2 \epsilon / 2$. \square

Since the partitions F and ClusMin are the same for all pairs where at least one element is in a large set of ClusMin, an immediate consequence is that the solution returned by the algorithm has cost at most $\gamma n^2(1 + \epsilon/3)(1 + \epsilon/2)$ which is at most equal to $\gamma n^2(1 + \epsilon)$ for any sufficiently small ϵ . The following technical result completes our proof by showing that Lemma 3.5 holds. The proof is a mechanical consequences of the values of β , $|T_{low}|$ and ϵ' .

Lemma 3.6. $l(n-1)|T_{low}| \left(2\beta + \frac{|T_{low}|}{l(n-1)}\right) \leq \frac{\epsilon}{2} \gamma n^2$.

Proof. Since $|T_{low}| \leq \frac{40\gamma nk^2}{l}$ and $\beta = \frac{\epsilon}{20 \cdot 160k^2 l}$, it suffices to prove that $l(n-1) \frac{40\gamma nk^2}{l} \left(\frac{\epsilon}{20 \cdot 80k^2 l} + \frac{40\gamma nk^2}{l^2(n-1)}\right) \leq \frac{\epsilon}{2} \gamma n^2$ that is equivalent to $\frac{80(n-1)k^2}{l} \left(\frac{\epsilon}{20 \cdot 80k^2} + \frac{40k^2 \gamma n}{l(n-1)}\right) \leq \epsilon n$. Since we are only interested in instances where the algorithm of [3] fails to provide a $(1 + \epsilon)$ approximation ratio, we can assume that $\gamma < \epsilon' = \frac{\epsilon}{128 \cdot 20^2 k^4}$, consequently it suffices to prove that $\frac{80(n-1)k^2}{l} \left(\frac{\epsilon}{20 \cdot 80k^2} + \frac{2 \cdot 20k^2 \epsilon n}{128 \cdot 20^2 l(n-1)k^4}\right) \leq \epsilon n$ that is equivalent to $\frac{4(n-1)}{l} \left(\frac{1}{80} + \frac{n}{64l(n-1)}\right) \leq n$ which in turn is equivalent to $9n \leq 80ln + 4n$ which is trivially true. \square

To complete the section and the analysis of the algorithm, we need to prove that the assumptions that we have made in some of the previous lemmas actually hold. The proofs are mechanical and quite tedious consequences of the values of β , γ and ϵ' .

Lemma 3.7. If $n \geq 16k^2$ then $\frac{n}{2k} - \gamma n^2 \frac{40k^2}{l(n-1)} > 2(k+1) \left(\frac{1}{20k^2} + \beta\right)n + 1$.

Proof. By the values of γ and β , and since we can assume that $\gamma < \epsilon' = \frac{\epsilon}{128 \cdot 20^2 k^4}$, the inequality can be rewritten as $\frac{n}{2k} - \frac{\epsilon}{128 \cdot 20^2 k^4} n^2 \frac{40k^2}{l(n-1)} > 2(k+1)$

1) $((\frac{1}{20k^2} + \frac{\epsilon}{128 \cdot 20^2 k^4})n + 1)$ which can be simplified as $\frac{n}{2k} - \frac{\epsilon}{64 \cdot 20k^2} n^2 \frac{1}{l(n-1)} > 2(k+1)((\frac{1}{20k^2} + \frac{\epsilon}{128 \cdot 20^2 k^4})n + 1)$. Since $\frac{n}{n-1} \leq 2$, it suffices to prove that $\frac{n}{2k} (1 - \frac{\epsilon}{16 \cdot 20kl}) > 2(k+1)((\frac{1}{20k^2} + \frac{\epsilon}{128 \cdot 20^2 k^4})n + 1)$. As $k, l \geq 2$ and ϵ is tiny, $\frac{\epsilon}{16 \cdot 20kl} < \frac{1}{1000}$, therefore we are only interested in proving that $\frac{999}{1000} \cdot \frac{n}{2k} > 2(k+1)((\frac{1}{20k^2} + \frac{\epsilon}{128 \cdot 20^2 k^4})n + 1)$ which is equivalent to $\frac{999}{1000} \cdot \frac{n}{2k} > (k+1)(\frac{1}{10k^2} + \frac{\epsilon}{64 \cdot 20^2 k^4})n + 2(k+1)$. Again, $\frac{k+1}{k} \leq 2$, therefore it is sufficient to prove that $\frac{999}{1000} \cdot \frac{n}{2k} > (\frac{1}{5k} + \frac{\epsilon}{32 \cdot 20^2 k^3})n + 2(k+1)$ which is equivalent to $\frac{599}{1000} \cdot \frac{n}{2k} > \frac{\epsilon}{32 \cdot 20^2 k^3} n + 2(k+1)$. Since $k \geq 2$, $\frac{\epsilon}{32 \cdot 20^2 k^3} < \frac{1}{1000}$, hence it suffices to prove that $\frac{n}{4k} > 2(k+1)$ which is an immediate consequence of the assumption $n \geq 16k^2$. \square

Lemma 3.8. *If $n \geq 16k^2$ then $2(\frac{1}{20k^2} + \beta)n + k < \frac{n}{2k} - \frac{80\gamma nk^2}{l}$.*

Proof. By the values of γ and β the inequality can be rewritten as $2(\frac{1}{20k^2} + \frac{\epsilon}{128 \cdot 20^2 k^4})n + k < \frac{n}{2k} - \frac{80nk^2}{l} \frac{\epsilon}{128 \cdot 20^2 k^4}$ which can be simplified as $\frac{n}{k} (\frac{1}{5k} + \frac{\epsilon}{32 \cdot 20^2 k^3} + \frac{\epsilon}{16 \cdot 20lk}) + 2k < \frac{n}{k}$. As $k, l \geq 2$, it is immediate to notice that $\frac{1}{5k} + \frac{\epsilon}{32 \cdot 20^2 k^3} + \frac{\epsilon}{16 \cdot 20lk} \leq \frac{1}{4}$, therefore it suffices to prove that $2k < \frac{3n}{4k}$, which is an immediate consequence of the assumption $n \geq 16k^2$. \square

4 NP-hardness

In this section we prove that 2-MIN-CC(3) is NP-hard. From the NP-hardness of 2-MIN-CC, it is easy to show that also k -MIN-CC(3) is NP-hard for any fixed k . Our proof consists of a reduction from the NP-hard Min Bisection Problem (MIN-BIS) to 2-MIN-CC(3). The MIN-BIS problem, given a graph $G = (V, E)$, asks for a partitioning of V in two equal-sized sets so that the number of edges connecting vertices in different sets is minimized.

For our purposes, in this section we give a different, but equivalent, definition of cost of a solution π of MINIMUM CONSENSUS CLUSTERING over instance Π can be alternatively defined as:

$$\sum_{\forall(i < j)} (r_\pi(i, j)d_\Pi(i, j) + (1 - r_\pi(i, j))s_\Pi(i, j)), \quad (1)$$

where $r_\pi(i, j) = 1$ iff (i, j) are co-clustered in π , otherwise $r_\pi(i, j) = 0$. The above formula will be used in the paper (see Section 4) to define the cost of a set P of pairs in a solution π as $\sum_{\forall(i, j) \in P} (r_\pi(i, j)d_\Pi(i, j) + (1 - r_\pi(i, j))s_\Pi(i, j))$.

Given an instance $G = (V, E)$ of MIN-BIS, where $|V| = n$ and $|E| = m$, we build an instance of 2-MIN-CC(3) as follows.

First we define the universe set V . For each $v_i \in V$, we define a set of n^4 elements $X_i = \{x_{i,1}, \dots, x_{i,n^4}\}$, and a set of n elements $Y_i = \{y_{i,1}, \dots, y_{i,n}\}$. The universe set is $V = (\cup_i X_i \cup Y_i)$. Next we define the three input partitions of 2-MIN-CC(3), $\Pi = \{\pi_1, \pi_2, \pi_3\}$. Partitions π_1 and π_2 are identical and consist of n disjoint sets $X_i \cup Y_i$, with $i = 1, \dots, n$. The partition π_3 contains the sets X_i , moreover for each edge $(v_i, v_j) \in E$, in π_3 we have the set $\{y_{i,h}, y_{j,l}\}$ consisting of two elements taken respectively from Y_i and Y_j (the actual elements taken are not important, provided that π_3 is a partition of the universe set – which is trivial to obtain). Finally, in π_3 we have a singleton for each element of $\cup Y_i$ that are not in a two-element set according to the previous rule.

Observation 4.1. *Since all the elements in X_i are co-clustered in all input partitions, each X_i is contained in a set of the optimal solution.*

The previous observation allows ourselves to restrict our attention to solutions where all elements of X_i are co-clustered. Consider a solution $\pi = (S_1, S_2)$. The cost of π can be expressed as the cost of all pairs of elements in π . We can split the cost of π into four parts:

1. the cost of pairs of elements both belonging to $\cup X_i$,
2. the cost of pairs of elements with exactly one element belonging to $\cup X_i$,
3. the cost of pairs of elements in $Y_i \times Y_j$ with $i \neq j$,
4. the cost of pairs of elements both belonging to a set Y_i .

We will call *balanced* a solution (S_1, S_2) where both S_1 and S_2 contain exactly $\frac{n}{2}$ sets X_i . The following lemma states that optimal solutions must be balanced.

Lemma 4.2. *Let $\pi = (S_1, S_2)$ be a solution of 2-MIN-CC(3), then the cost of π is at most $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + 3n^7 + \frac{3}{2}n^4 - \frac{3}{2}n^3$ if and only if π is a balanced solution.*

Proof. Notice that the total cost of case 2) is at most $3n^2 \cdot n^5 = 3n^7$ as $|\cup Y_i| = n^2$ and $|\cup X_i| = n^5$, while the sum of total costs of cases 3) and 4) is at most $3\binom{n^2}{2} = \frac{3}{2}n^4 - \frac{3}{2}n^3$.

Let z be the number of sets X_i included in S_1 . The cost of the pairs of elements both belonging to $\cup X_i$ is $C(z) = 3 \left(\binom{z}{2} + \binom{n-z}{2} \right) n^8$. Indeed, only the pairs of elements in distinct sets X_i that are co-clustered in S_1 and S_2 contribute to the cost, as no pair of elements belonging to two distinct sets X_i is co-clustered in an input partition. The minimum of $C(z)$ is attained for $z = \frac{n}{2}$. For any other z , the value of $C(z)$ is at least equal $C(\frac{n}{2} - 1)$.

Since $C(\frac{n}{2}) = \frac{3}{4}n^{10} - \frac{3}{2}n^9$, the maximum total cost for a balanced solution is $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + 3n^7 + \frac{3}{2}n^4 - \frac{3}{2}n^3$, while the maximum total cost for an unbalanced solution is at least $C(\frac{n}{2} - 1) = \frac{3}{4}n^{10} - \frac{3}{2}n^9 + 3n^8 > \frac{3}{4}n^{10} - \frac{3}{2}n^9 + 3n^7 + \frac{3}{2}n^4 - \frac{3}{2}n^3$. \square

From Lemma 4.2 we can consider only balanced solutions. A balanced solution π is called *standard* if, for each i , X_i and Y_i are contained in the same set of π . The following lemma shows that we can consider only standard solutions

Lemma 4.3. *Let $\pi = (S_1, S_2)$ be a balanced solution of 2-MIN-CC(3), then the cost of π is at most $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{1}{2}n^3 - \frac{1}{2}n^2$ iff π is a standard solution.*

Proof. Let $\pi = (S_1, S_2)$ be a balanced solution, then the total cost of pairs of elements with exactly one element belonging to $\cup X_i$ is at most $\frac{3}{4}n^7 - \frac{1}{2}n^6$ as all pairs in $X_i \times Y_j$, with $i \neq j$, contribute with a cost 3 if and only if $X_i \cup Y_j$ is contained in a set of π , and have no cost otherwise. At the same time all pairs in $X_i \times Y_i$ have cost 1 in any standard solution, as $X_i \cup Y_i$ are a set of two input partitions, while in the third input partition, π_3 , no pairs in $X_i \times Y_i$ are co-clustered. If π is a standard solution, then the total cost of pairs of elements in $Y_i \times Y_j$ with $i \neq j$ is $\frac{1}{4}n^4$ as only half of such pairs are co-clustered in a standard solution. Following the reasoning of the proof of Lemma 4.2, with our new estimates of cases 2) and 3), it is immediate to notice that, if π is a standard solution, then its cost is at most $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + n\binom{n}{2} = \frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{1}{2}n^3 - \frac{1}{2}n^2$.

Now assume that π is not a standard solution, that is there exists an element $y \in Y_i$ that is not clustered together with all elements of X_i . Again, following the same lines of the proof of Lemma 4.2, the cost of π is at least $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + n^4$, as all pairs in $\{y\} \times X_i$ have a cost 2, instead of 1 as in a standard partition. Since $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + n^4 > \frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{1}{2}n^3 - \frac{1}{2}n^2$, the lemma follows. \square

Given a standard solution π , by construction of the reduction, with each edge $(v_i, v_j) \in E$, we associate a pair $\{y_{i,h}, y_{j,l}\}$. Let us denote by F the set

of such pairs, and by F_c the subset of all pairs in F that are co-clustered in π . We conclude the proof with the following theorem.

Theorem 4.4. *Let $G = (V, E)$ be an instance of MIN-BIS, and let (π_1, π_2, π_3) be its associated instance of 2-MIN-CC(3). Then (π_1, π_2, π_3) has a solution of cost $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + 3/4n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{3}{2}n^4 - \frac{1}{2}n^3 - \frac{1}{2}n^2 + (|F| - k) - k$ if and only if G has a bisection of cost k .*

Proof. Let (V_1, V_2) be a bisection with cost k . Then let S_1 be the set $\cup_{i \in V_1} (X_i \cup Y_i)$, and let $S_2 = \cup_{i \in V_2} (X_i \cup Y_i)$. By construction (S_1, S_2) has cost $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{3}{2}n^4 - \frac{1}{2}n^3 - \frac{1}{2}n^2 + (|F| - k) - k$.

Now let (S_1, S_2) be a solution of 2-MIN-CC(3) with cost $\frac{3}{4}n^{10}/4 - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{3}{2}n^4 - \frac{1}{2}n^3 - \frac{1}{2}n^2 + (|F| - k) - k$. By Lemmas 4.2, 4.3 (S_1, S_2) must be a standard solution.

Recall that the cost of a solution $\pi = (S_1, S_2)$ can be expressed as the cost of all pairs of elements in π , such a cost can be split into parts 1), 2), 3) and 4). Moreover, following the proof of Lemmas 4.2, 4.3, we know that the total cost of case 1) is $\frac{3}{4}n^{10} - \frac{3}{2}n^9$, the total cost of case 2) is $\frac{3}{4}n^7 - \frac{1}{2}n^6$. By direct inspection the total cost of case 4) is $\frac{1}{2}n^3 + \frac{1}{2}n^2$.

We still have to consider case 3), that is the cost of pairs $(y_{i,q}, y_{j,t})$, with $j \neq i$. We have to distinguish three cases, according to the fact that $(y_{i,q}, y_{j,t}) \in F - F_c$ (in this case the cost is 1), $(y_{i,q}, y_{j,t}) \in F_c$ (in this case the cost is 2), $(y_{i,q}, y_{j,t}) \notin F$ (in this case the cost is 3 if $y_{i,q}$ and $y_{j,t}$ are co-clustered, and 0 otherwise). Therefore the total cost of case 3) can be written as $n^2 \binom{n}{2} + |F - F_c| - |F_c|$.

Summing up the costs of the four cases we obtain a total cost $\frac{3}{4}n^{10} - \frac{3}{2}n^9 + \frac{3}{4}n^7 - \frac{1}{2}n^6 + \frac{1}{4}n^4 + \frac{3}{2}n^4 - \frac{1}{2}n^3 - \frac{1}{2}n^2 + |F| - 2|F_c|$. Consequently, taking into account the initial hypothesis, $|F_c| = k$. Let (V_1, V_2) be the solution of G where $V_1 = \{v_i | X_i \subseteq S_1\}$ and $V_2 = V - V_1$. By construction the number of edges of E crossing the bipartition (V_1, V_2) is equal to $|F_c|$ which, in turn, is equal to k completing the proof. \square

5 Conclusions

In this paper we have studied the MINIMUM CONSENSUS CLUSTERING problem when the output partition contains at most a constant number of sets. We have shown that the MinDisAg algorithm [7] can be applied also for our problem, hence showing that its applicability is not restricted to unweighted problems. Moreover we have proved that the same problem is NP-hard even on instances of three input partitions, thereby justifying our reliance on polynomial time approximation algorithms.

In our opinion the main idea behind MinDisAg algorithm could be applied to some more general versions of both MINIMUM CONSENSUS CLUSTERING and MINIMUM CORRELATION CLUSTERING than the ones studied here and in [7].

Acknowledgments

PB, and GDV have been partially supported by FAR 2008 grant “Computational models for phylogenetic analysis of gene variations”. PB has been partially supported by the MIUR PRIN 2007 Project “Mathematical aspects and emerging applications of automata and formal languages”.

References

- [1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5), 2008.
- [2] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.
- [3] P. Bonizzoni, G. Della Vedova, R. Dondi, and T. Jiang. On the approximation of correlation clustering and consensus clustering. *J. Comput. Syst. Sci.*, 74(5):671–696, 2008.
- [4] M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. *J. Comput. Syst. Sci.*, 71(3):360–383, 2005.
- [5] V. Filkov and S. Skiena. Heterogeneous data integration with the consensus clustering formalism. In *Data Integration in the Life Sciences, First International Workshop, (DILS)*, pages 110–123, 2004.
- [6] V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools*, 13(4):863–880, 2004.
- [7] I. Giotis and V. Guruswami. Correlation clustering with a fixed number of clusters. *Theory of Computing*, 2(13):249–266, 2006.
- [8] M. Grötschel and Y. Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45:52–96, 1989.

- [9] M. Krivanek and J. Moravek. Hard problems in hierarchical-tree clustering. *Acta Informatica*, 23:311–323, 1986.
- [10] C. Swamy. Correlation clustering: maximizing agreements via semidefinite programming. In *Proc. 15th Symp. on Discrete Algorithms (SODA)*, pages 526–527, 2004.
- [11] Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas*, 3(3):323–349, 1998.